**CROP RECOMMENDATION SYSTEM**

CAPSTONE PROJECT

Submitted in partial fulfillment of the requirements of the

**Post Graduate Certification Program**

**in**

**Artificial Intelligence and Machine Learning**

By

**IMTIYAZ ALAMSHAH 2024AIML132**

**TEJASWINI LALITHA 2024AIML587**

**ANKIT MISHRA 2024AIML509**

**BAGEPALLI SOUMYASREE 2024AIML606**

Under the supervision of

**Prof. Aniruddha Dasgupta**

Project work carried out at

**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE**

**Pilani (Rajasthan) INDIA**

(October,2025)

**PGCP AIML CAPSTONE PROJECT**

**CROP RECOMMENDATION SYSTEM**

Submitted in partial fulfillment of the requirements of the

PGCP - Artificial Intelligence and Machine Learning

By

**IMTIAYZ ALAMSHAH 2024AIML132**

**LIPSA MISHRA 2024AIML067**

**RAIBAGKAR HARSHAL GOPALRAO 2024AIML057**

**SHAH HIMA KIRANBHAI 2024AIML011**

**UPADYAYULA V SHARMA 2024AIML070**

Under the supervision of

**Prof. Aniruddha Dasgupta**

Project work carried out at

**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE**

**PILANI (RAJASTHAN)**

(October, 2025)

**ACKNOWLEDGEMENTS**

We express our deep sense of gratitude to BITS, Pilani for providing an opportunity to pursue our project on “**Crop Recommendation System**”.

We thank our mentor, **Prof. Aniruddha Dasgupta** for the constant support and guidance which resulted in the successful completion of the project within the specified time. His unflinching help and encouragement were a constant source of inspiration to us.

We would also like to thank all our professors, who taught us the basic and advanced concepts of artificial intelligence and machine learning in such a nuanced manner, during this certification program.

We would also like to thank our respective families and organizations for supporting us through this journey.

**IMTIAYZ ALAMSHAH (2024AIML132)**

**LIPSA MISHRA (2024AIML067)**

**RAIBAGKAR HARSHAL GOPALRAO (2024AIML057)**

**SHAH HIMA KIRANBHAI (2024AIML011)**

**UPADYAYULA V SHARMA (2024AIML070)**

**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI**

**CERTIFICATE**

This is to certify that the Capstone Project entitled “**Crop Recommendation System**”

and submitted by **IMTIAYZ ALAMSHAH (2024AIML132), LIPSA MISHRA (2024AIML067), RAIBAGKAR HARSHAL GOPALRAO (2024AIML057), SHAH HIMA KIRANBHAI (2024AIML011) AND UPADYAYULA V SHARMA (2024AIML070)** in partial fulfillment of the requirements of PGCP AIML Capstone Project, embodies the work done by him/her under my supervision.

Place : Hyderabad Signature of the Mentor

Date : 20-10-2025 Name : Prof. Aniruddha Dasgupta

**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI**

**2024-25**

**PGCP AIML CAPSTONE PROJECT**

Project Title : Crop Recommendation System

Name of Mentor : Prof. Aniruddha Dasgupta

Name of Student : IMTIAYZ ALAMSHAH, LIPSA MISHRA, RAIBAGKAR HARSHAL GOPALRAO, SHAH HIMA KIRANBHAI, UPADYAYULA V SHARMA

ID No. of Student : 2024AIML132, 2024AIML067, 2024AIML057, 2024AIML011, 2024AIML070

## **ABSTRACT**

Agriculture is the backbone of India’s economy, supporting a majority of the rural population. However, farmers often face challenges in selecting the most suitable crop due to variations in soil type, climatic conditions, and lack of technical knowledge. The **Crop Recommendation System in India** addresses this issue by providing a scientific and data-driven approach to crop selection. It leverages agricultural and environmental data to assist farmers in making informed decisions that can enhance productivity and sustainability.

The system uses machine learning algorithms to analyze parameters such as soil type, pH level, nutrient composition, temperature, and rainfall. Based on these inputs, it predicts the most appropriate crop for a specific region. By integrating technology with agriculture, the project promotes efficient resource utilization, improved crop yield, and long-term food security. This initiative demonstrates the potential of artificial intelligence in revolutionizing traditional farming practices in India.

.

## **TABLE OF CONTENTS**

[Problem Statement 7](#_ik3p3tyg8xd9)

[Objective of the Project 7](#_gmge5zlth7bj)

[Machine Learning Process Flow (Architecture) 8](#_vjy1qobjakpm)

[Resources or Prerequisite 8](#_9u20vf5f4fyv)

[Potential data challenges & risks in doing the project 9](#_2079qjstd4m0)

[Detailed Plan of Work 10](#_fvx49uizcv5i)

[Dataset 11](#_3pruo97ks3r3)

[Pre-Processing and Feature Engineering 12](#_wb5ox9kngkzj)

[Machine Learning Modeling & Techniques Applied 17](#_etd7qhipll84)

Objective: 17

Modeling Approach: 17

[Models Trained: 17](#_izgnf2gyux9y)

[Training Details: 17](#_td2k1y3ehutk)

[Code & Screenshots 18](#_ejbim3i30hi8)

[Deployment 28](#_x1q8dzt9fbsu)

Objective: 28

Tools and Technologies Used: 28

[Deployment Architecture Overview: 28](#_7qhdh7m5aeb7)

[Project Folder Structure: 28](#_p0n3u9sotzmi)

[Key Insights & Impactful/Important Features 29](#_f21jv3aly7ut)

[Key Insights 29](#_bx96rvci6o30)

[Important Features Identified 29](#_9xtqsvfrfkxm)

[Future Work / Scope of improvements 30](#_s1wqmwd1efjl)

[1. Expansion of Historical Data 30](#_a2cqskye1s6m)

[2. Incorporating More Features 30](#_1mwdb23hez6x)

[BIBLIOGRAPHY 31](#_10gvsajen0il)

[Checklist of items for the Final report 31](#_jgnjez5pipxl)

## 

## **Problem Statement**

Agriculture in India is highly dependent on environmental conditions and soil characteristics, which vary widely across regions. Most farmers rely on traditional knowledge or past experience to decide which crop to cultivate, often without access to scientific data or expert guidance. This leads to suboptimal crop choices, reduced productivity, and financial losses.

Unpredictable weather patterns, irregular rainfall, and soil degradation further complicate the decision-making process. Farmers are frequently unaware of how soil nutrients, pH level, or moisture content affect crop suitability. Consequently, they may invest time and resources into crops that are not well-suited to their land or local climate.

Therefore, there is a pressing need for an intelligent, data-driven system that can recommend the most suitable crop for a given location based on real-time agricultural parameters. Such a system can help optimize yields, ensure sustainable farming, and support better livelihood outcomes for Indian farmers.

## **Objective of the Project**

## The objective of the **Crop Recommendation System in India** is to design and implement an intelligent, data-driven application capable of predicting the most suitable crop for a given region using machine learning and environmental analysis. The project focuses on building a technically robust solution that processes agricultural data efficiently and provides accurate, actionable insights to farmers.

## **Odbjectives include:**

## To collect and preprocess agricultural datasets containing soil, weather, and crop parameters.

## To perform data cleaning, normalization, and feature extraction for model training.

## To implement and compare machine learning algorithms such as Random Forest, Decision Tree, and Support Vector Machine for crop prediction.

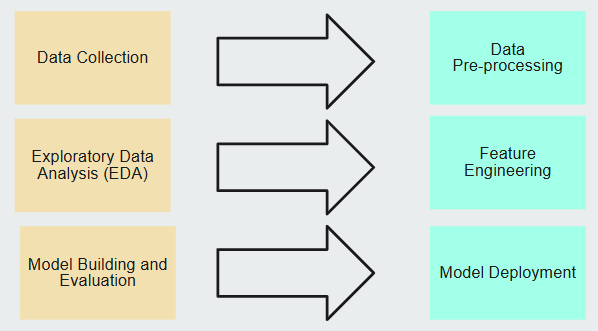
## To evaluate model accuracy using appropriate metrics like precision, recall, and F1-score.

## To develop a web or desktop-based user interface for inputting soil and climate data.

## To integrate the trained model with the interface to generate real-time crop recommendations.

## To deploy the system for testing and validate its performance using real-world agricultural data.

## **Machine Learning Process Flow (Architecture)**

****

## **Resources or Prerequisite**

Data Resources: Collect data from various sources <https://etrain.info/in>, <https://indianrail.gov.in>

Libraries & Frameworks:

* *pandas, numpy* for data manipulation
* *matplotlib, seaborn* for visualization
* *scikit-learn, xgboost* for machine learning
* *Streamlit* for deployment

Software: Python

## **Potential data challenges & risks in doing the project**

Developing a **Crop Recommendation System** involves handling complex and diverse agricultural datasets, which presents several data-related challenges and risks. The accuracy and effectiveness of the model depend heavily on the quality, completeness, and consistency of data used for training and testing.

One major challenge is the **availability and accessibility of reliable agricultural data**. Data on soil nutrients, weather conditions, and crop yields are often scattered across multiple government and private sources, making collection and integration difficult. Additionally, **data inconsistency and missing values** can impact the performance of machine learning algorithms, requiring extensive preprocessing and cleaning.

Another risk is **regional data imbalance**, where certain states or soil types are overrepresented, leading to biased predictions. **Temporal changes** in climate patterns and soil fertility can also make historical data less relevant over time. Finally, ensuring **data privacy and security** is critical, especially when integrating data from farmers or government databases. Addressing these challenges is essential to ensure the system’s accuracy, scalability, and long-term reliability.

## **Proposed Methodology**

The **Crop Recommendation System in India** aims to utilize data analytics and machine learning techniques to predict the most suitable crop for a given region based on soil and environmental parameters. The proposed methodology is divided into systematic stages that ensure efficient data handling, model training, and user interaction.

1. **Data Collection:**  
   Relevant datasets are collected from credible sources such as the data.gov.in, Kaggle data source, and government open data portals. The datasets include soil characteristics (pH, nitrogen, phosphorus, potassium), weather parameters (temperature, rainfall), and regional crop yield data.
2. **Data Preprocessing:**  
   Collected data undergo cleaning to remove missing, duplicate, or inconsistent entries. Feature scaling and normalization are applied to ensure uniformity. Outliers are identified and treated to improve model stability.
3. **Model Development:**  
   Various machine learning algorithms—such as **Random Forest**, **Decision Tree**, **Support Vector Machine (SVM)**, and **K-Nearest Neighbor (KNN)**—are trained and tested on the processed data. The models are tuned using hyperparameter optimization techniques to enhance prediction accuracy.
4. **Model Evaluation:**  
   Performance is assessed using metrics like accuracy, precision, recall, F1-score, and confusion matrix. Cross-validation techniques are applied to ensure the model’s generalization and reliability.
5. **Testing and Validation:**  
   The system is tested using real-time or region-specific data to validate the accuracy of recommendations. Feedback from agricultural experts and field trials can further refine the model.
6. **Deployment and Future Enhancement:**  
   After validation, the system is deployed for use by farmers and agricultural departments. Future enhancements may include IoT-based sensor integration, real-time weather APIs, and mobile application development for wider accessibility.

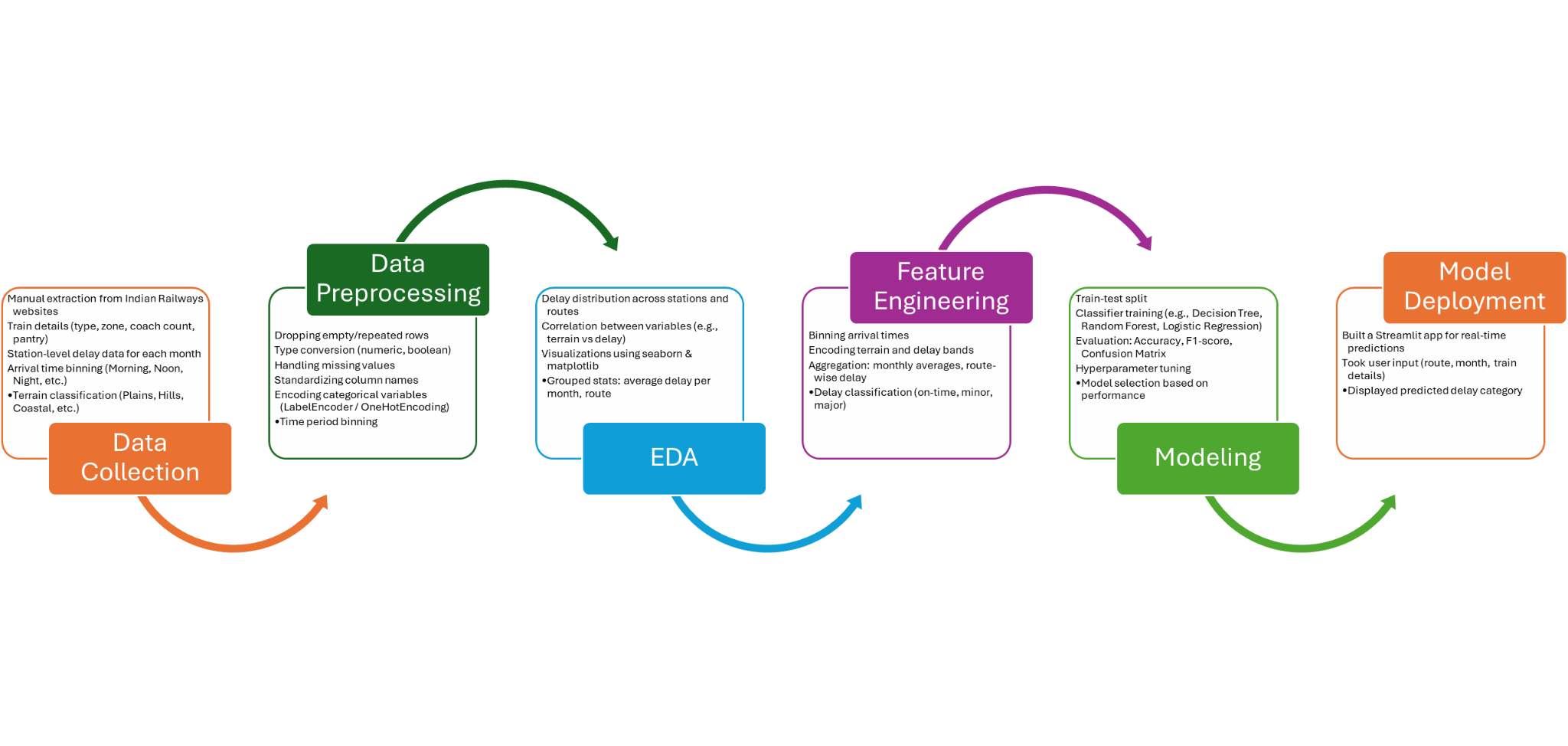
## **Detailed Plan of Work**

*Week 1-2: Data exploration and preprocessing.*

*Week 3-4: Feature engineering and EDA.*

*Week 5-6: Model training and hyperparameter tuning.*

*Week 7-8: Evaluation, deployment, and documentation*

****

**Problem Definition:** Identified the need to predict train delays (Low, Moderate, High) for trains arriving at Delhi from various cities to support operational planning.

**Data Collection:** Collected train and station data manually from [indianrail.gov.in](http://indianrail.gov.in) and [etrain.info](https://etrain.info/in) for selected routes (e.g., Bangalore–Delhi, Pune–Delhi).  
 Monthly delays were extracted by visually interpreting graphs and estimating (min, max, mode) delays.

**Data Preprocessing** Cleaned and standardized data, handled missing values, encoded categorical features (e.g., terrain, time slots).

**Exploratory Data Analysis (EDA)** Visualized delay patterns across train types, terrains, and time slots to identify key influencing factors.

**Feature Engineering** Created derived features such as terrain complexity, travel time, and monthly delay spread.

**Model Building and Evaluation** Trained multiple classification models (Random Forest, XGBoost, etc.).  
 Evaluated using accuracy, F1-score, and confusion matrix to select the best-performing model.

**Model Deployment** Deployed the final model using **Streamlit**, allowing users to input train details and get delay category predictions in real-time.

## **Dataset**

The dataset used in the Crop Recommendation System in India forms the foundation for training and evaluating the machine learning models. It consists of agricultural and environmental parameters that directly influence crop growth and yield. Each record in the dataset represents a unique set of soil and climatic conditions, along with the corresponding crop that thrives best under those conditions.

The dataset contains the following features (independent variables):

* **Nitrogen (N)**: Indicates the nitrogen content in the soil, essential for leaf growth and overall plant development.
* **Phosphorus (P)**: Represents the phosphorus level in the soil, important for root development and energy transfer.
* **Potassium (K)**: Measures the potassium content, crucial for water regulation, disease resistance, and fruit quality.
* **pH Value**: Denotes the acidity or alkalinity of the soil, which affects nutrient absorption and crop suitability.
* **Rainfall**: Average annual rainfall (in millimetres) recorded in the region, influencing water availability for crops.
* **Temperature**: The average temperature (in degrees Celsius), determining the thermal conditions suitable for different crops.

The **target variable** in the dataset is **Crop**, which specifies the name of the crop best suited for the given combination of soil nutrients, pH level, rainfall, and temperature. The dataset is typically labelled, meaning each data point has a known crop type associated with its environmental conditions.

This structured dataset enables supervised machine learning algorithms to learn the relationships between soil-climate parameters and optimal crop types. Once trained, the model can predict the most suitable crop for any new set of input parameters, making it a powerful decision-support tool for Indian farmers.

Total 40 types of Crop Labels are available in the data:

BlackPepper, BottleGourd, Brinjal, Cabbage, Cardamom, Cauliflower, Coriander, Cotton, Cucumber, Drumstick, Garlic, Grapes, HorseGram, JackFruit, Jowar, Jute, LadyFinger, Maize, Mango, Moong, Onion, Orange, Papaya, Pineapple, Potato, Pumpkin, Radish, Ragi, RapeSeed, Rice, SoyaBean, Sunflower, SweetPotato, Tomato, Turmeric and Wheat

## **Pre-Processing and Feature Engineering**

### **Feature Combination**

### Created new features by feature combinations:

* NPK\_Sum
* NP\_Ratio
* NK\_Ratio
* PK\_Ratio

### **Binning**

### Created new features by binning:

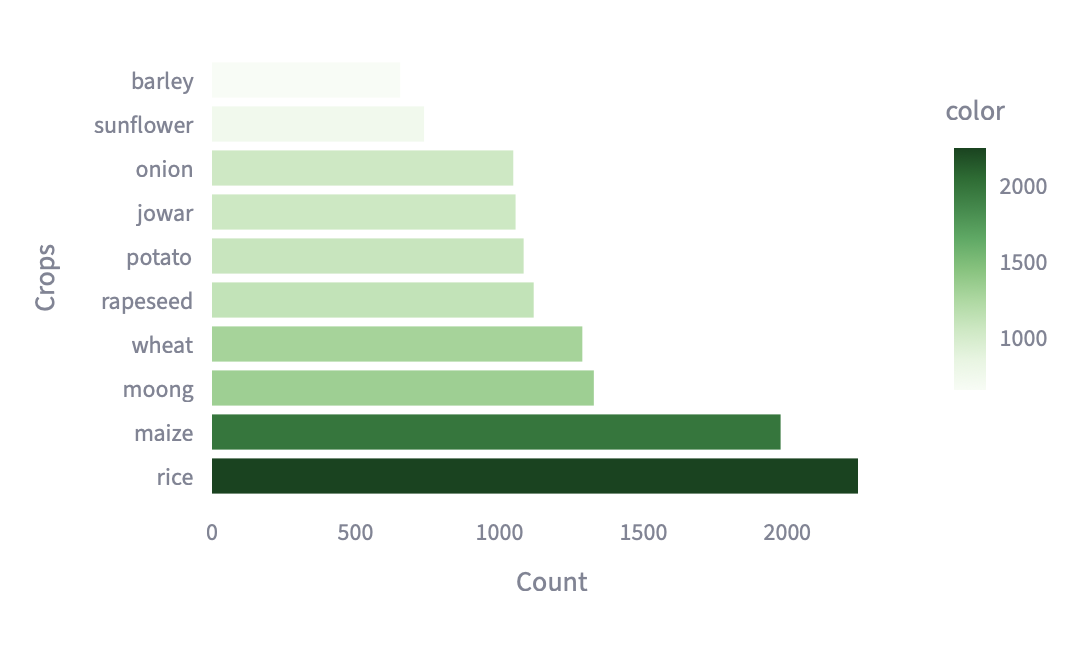
* ph\_acidic, ph\_netural, ph\_alkaline
* temp\_cold, temp\_moderate, temp\_warm, temp\_hot

# **Exploratory Data Analysis (EDA) Summary**

## **Overall Objectives**

The goal of the EDA was to understand:  
- The distribution of delays across trains and months  
- Variability in terrain, platform availability, and route length  
- Potential correlations between features like station count, coach type, and delay behavior

## **Top 10 crop distribution**

Feature 

**Feature Distribution**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

**Feature Correlation Matrix**

****

**Statistical Summary**



## **Machine Learning Modeling & Techniques Applied**

## **Objective:**

The **Crop Recommendation System in India** relies on machine learning techniques to predict the most suitable crop based on soil and environmental parameters such as nitrogen (N), phosphorus (P), potassium (K), pH value, rainfall, and temperature. The project adopts a **supervised learning approach**, where labeled data (features with corresponding crop types) is used to train predictive models capable of making accurate recommendations for unseen inputs.

## **Modeling Approach:**

We followed a consistent pipeline for all models:  
1. Preprocessed and engineered features using processed\_df  
2. Encoded categorical variables and binned numerical ones  
3. Used train\_test\_split() with stratification for fair model evaluation  
5. Evaluated using metrics like Accuracy, Confusion Matrix, and F1-score  
6. Saved each model and its corresponding input column structure and imputer (if needed)

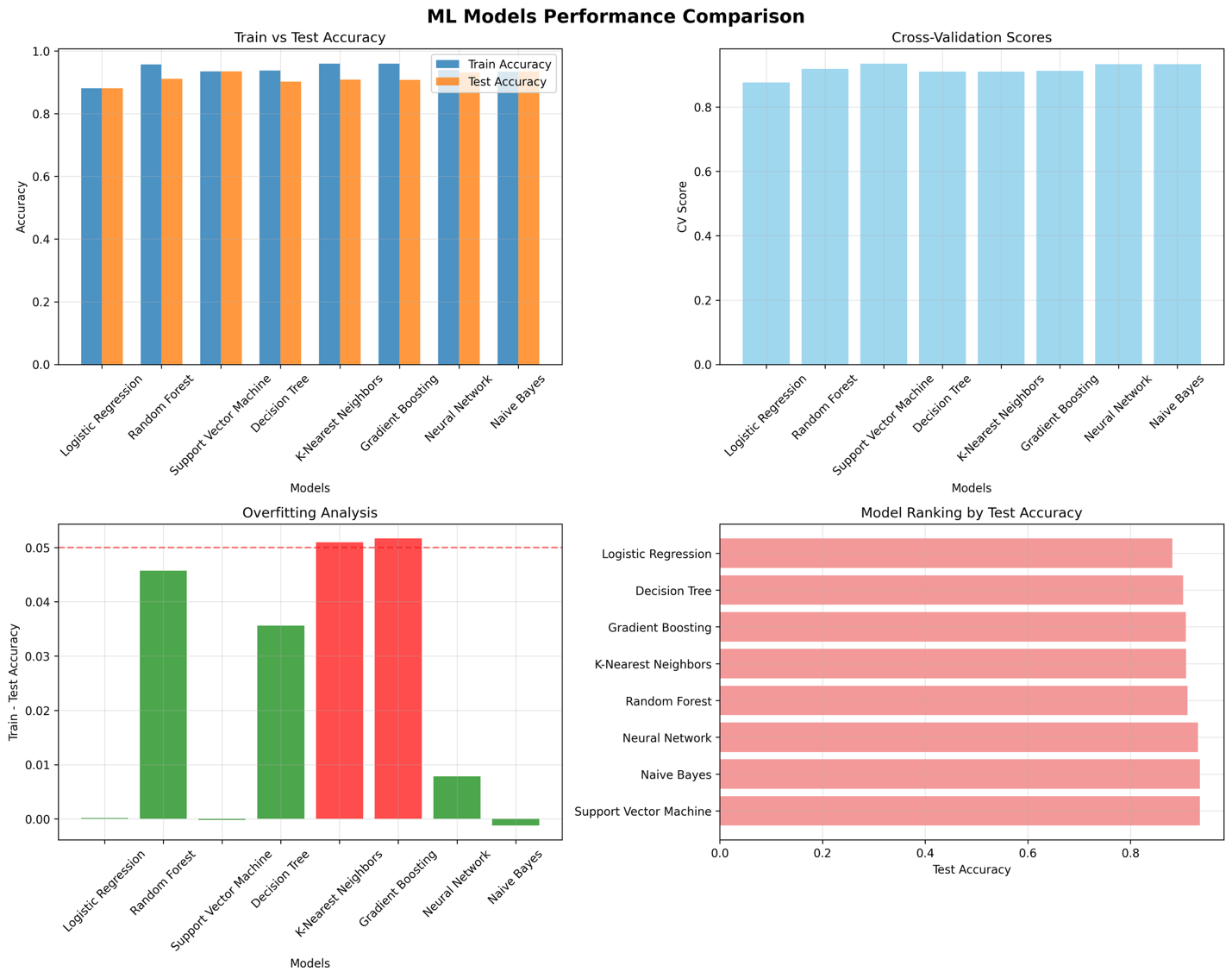
## 

## 

## **Models Trained:**

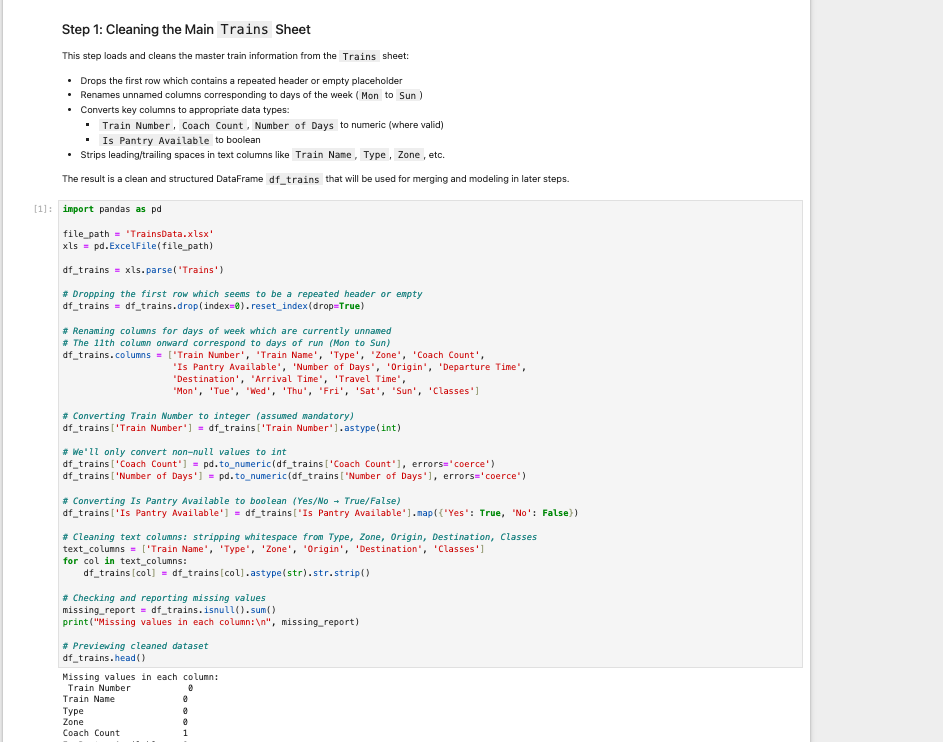
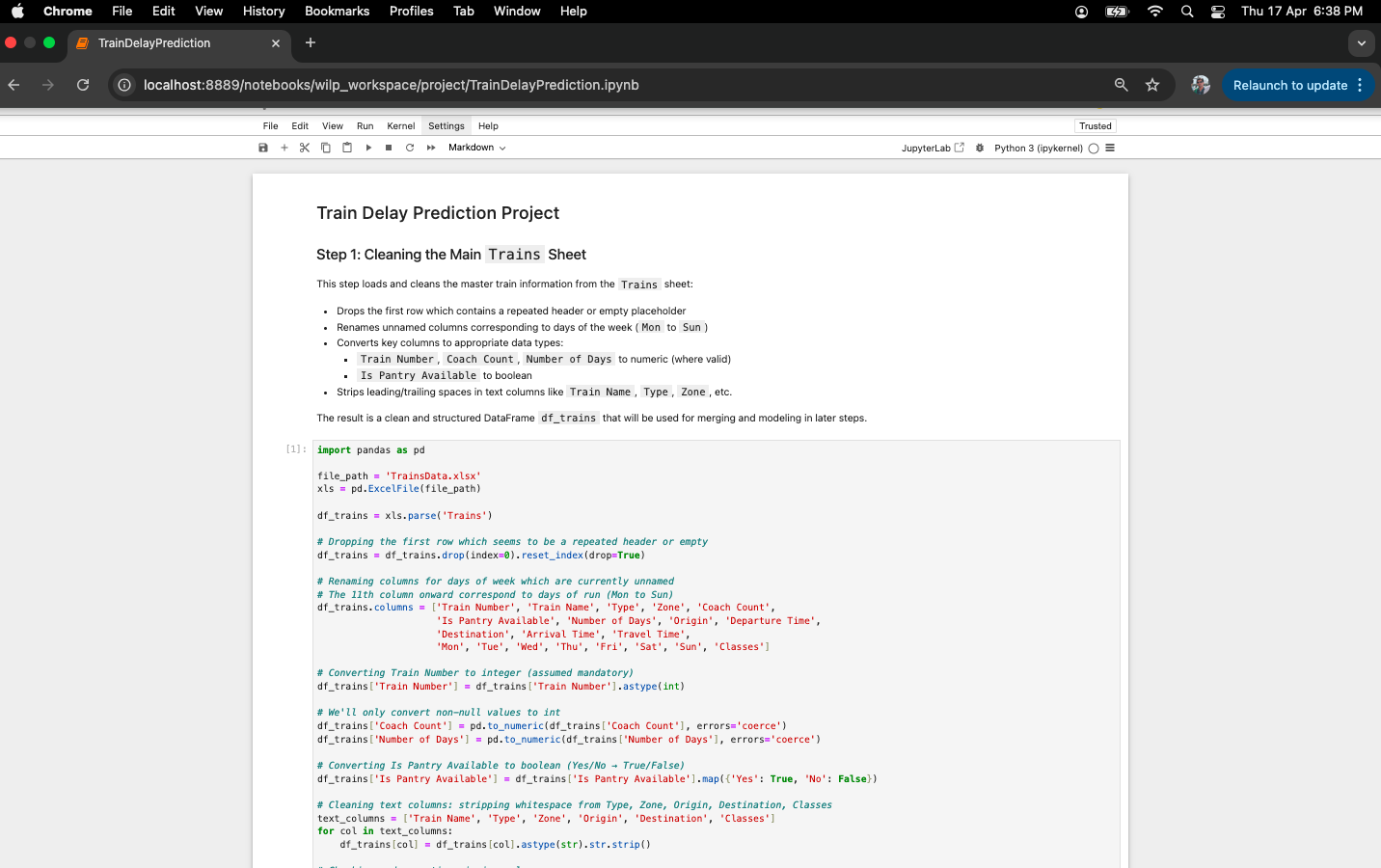
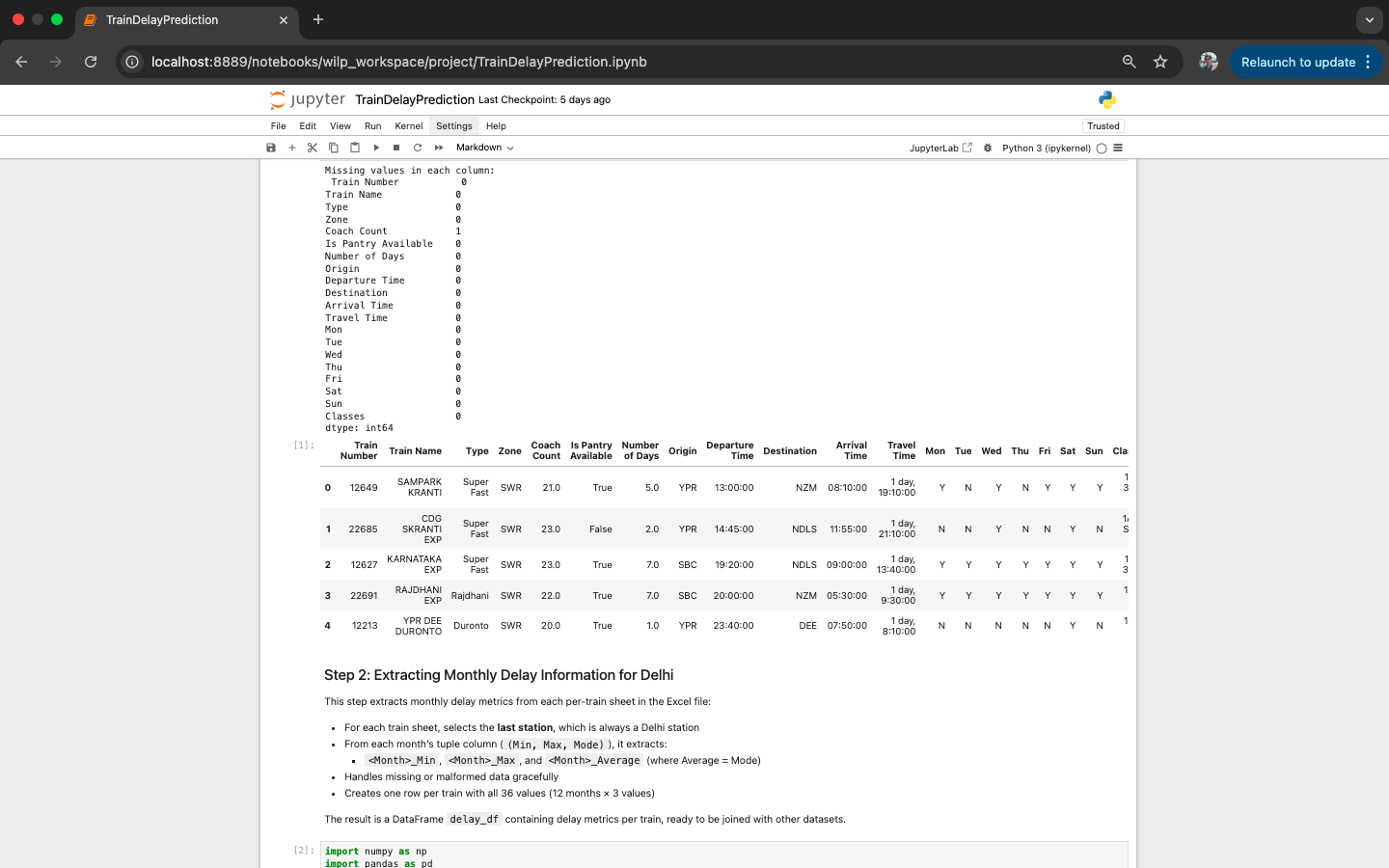
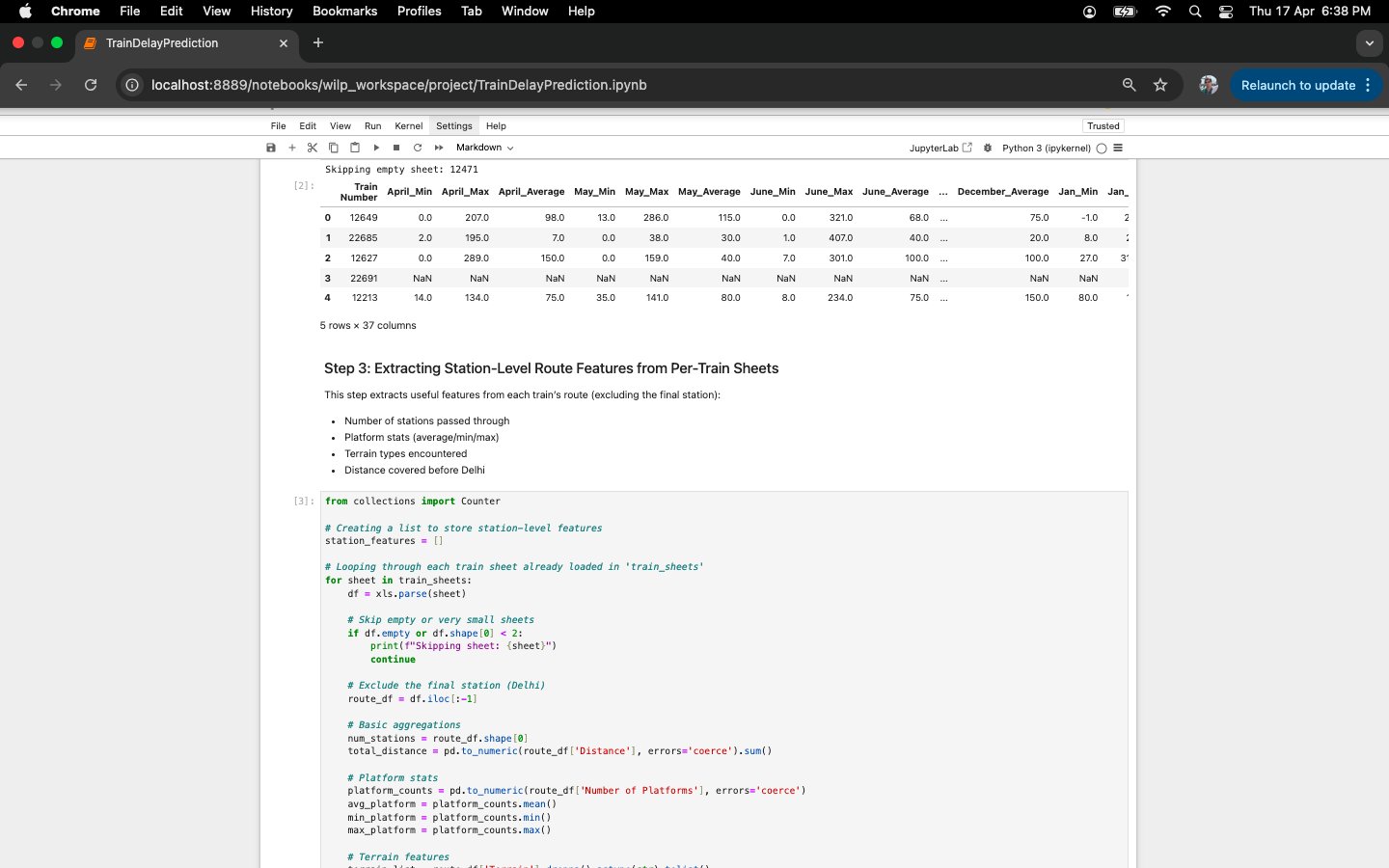
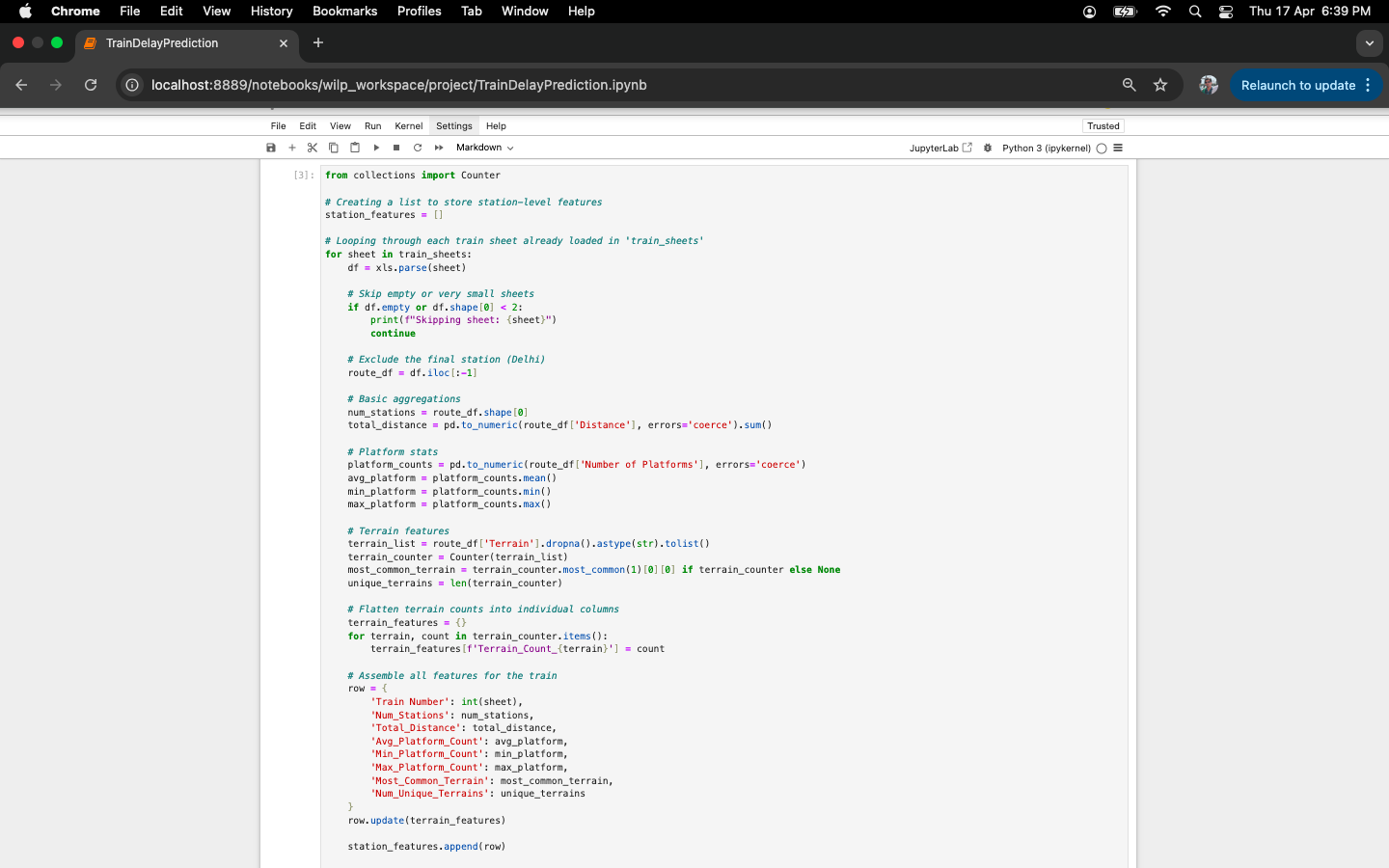
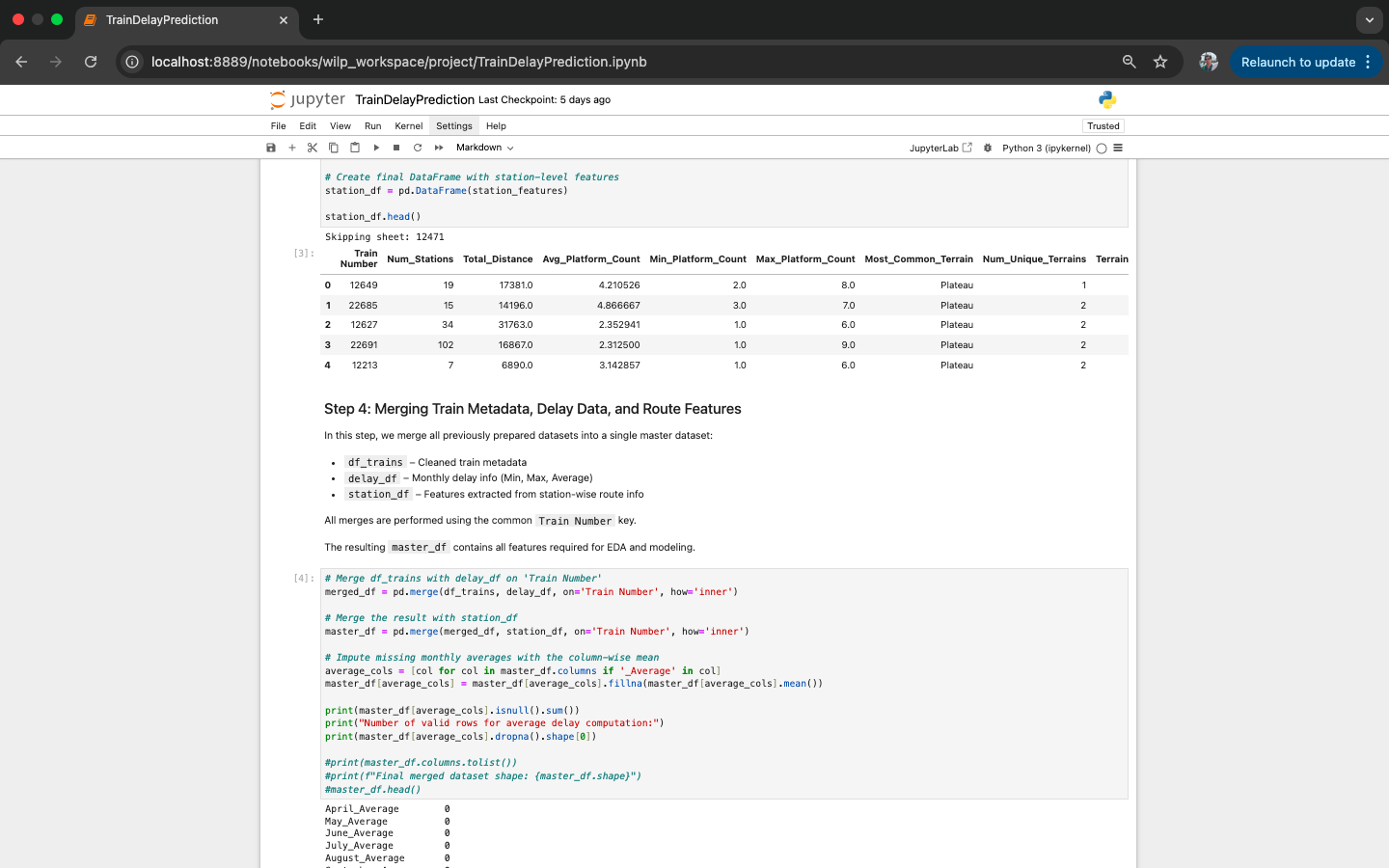
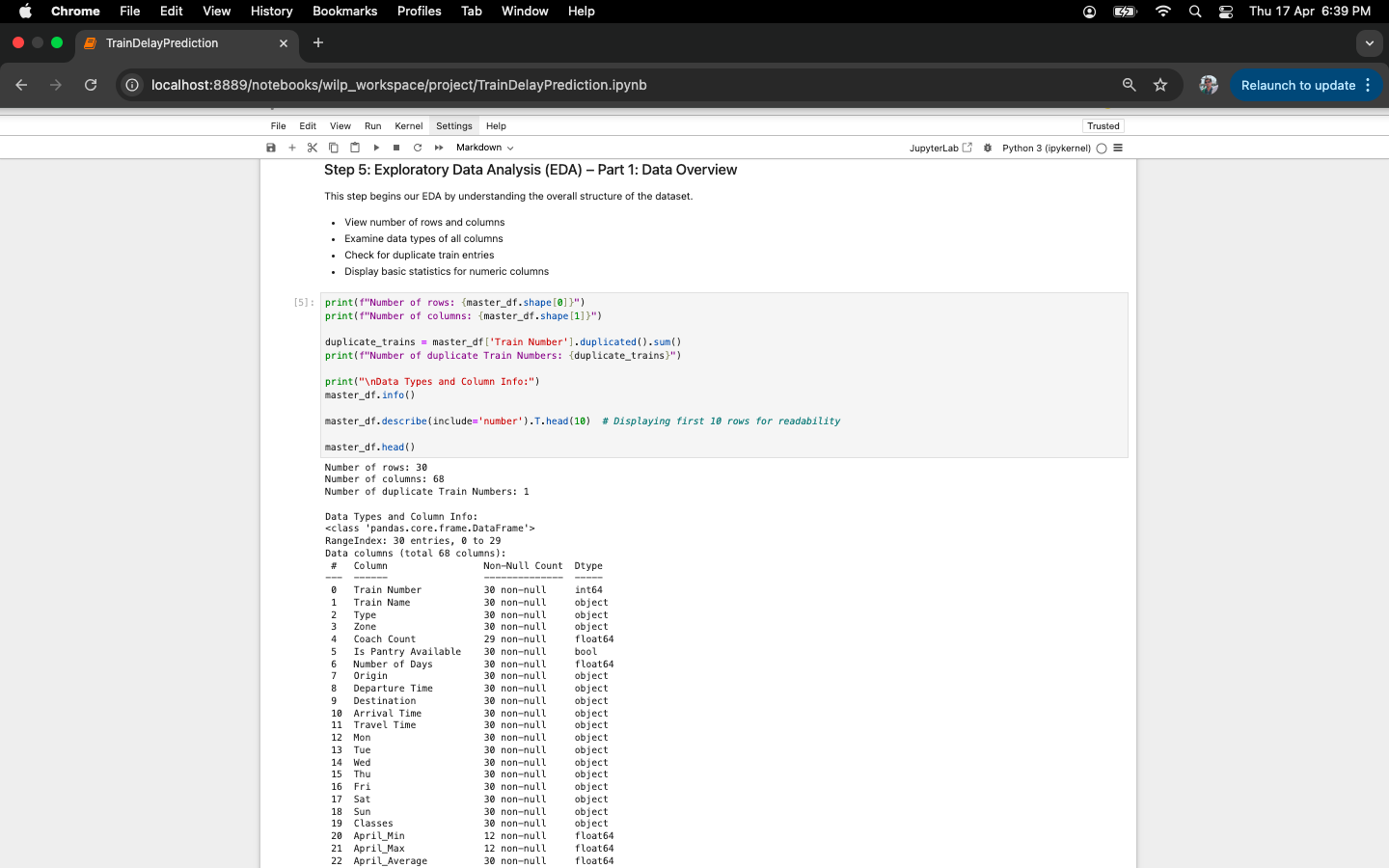
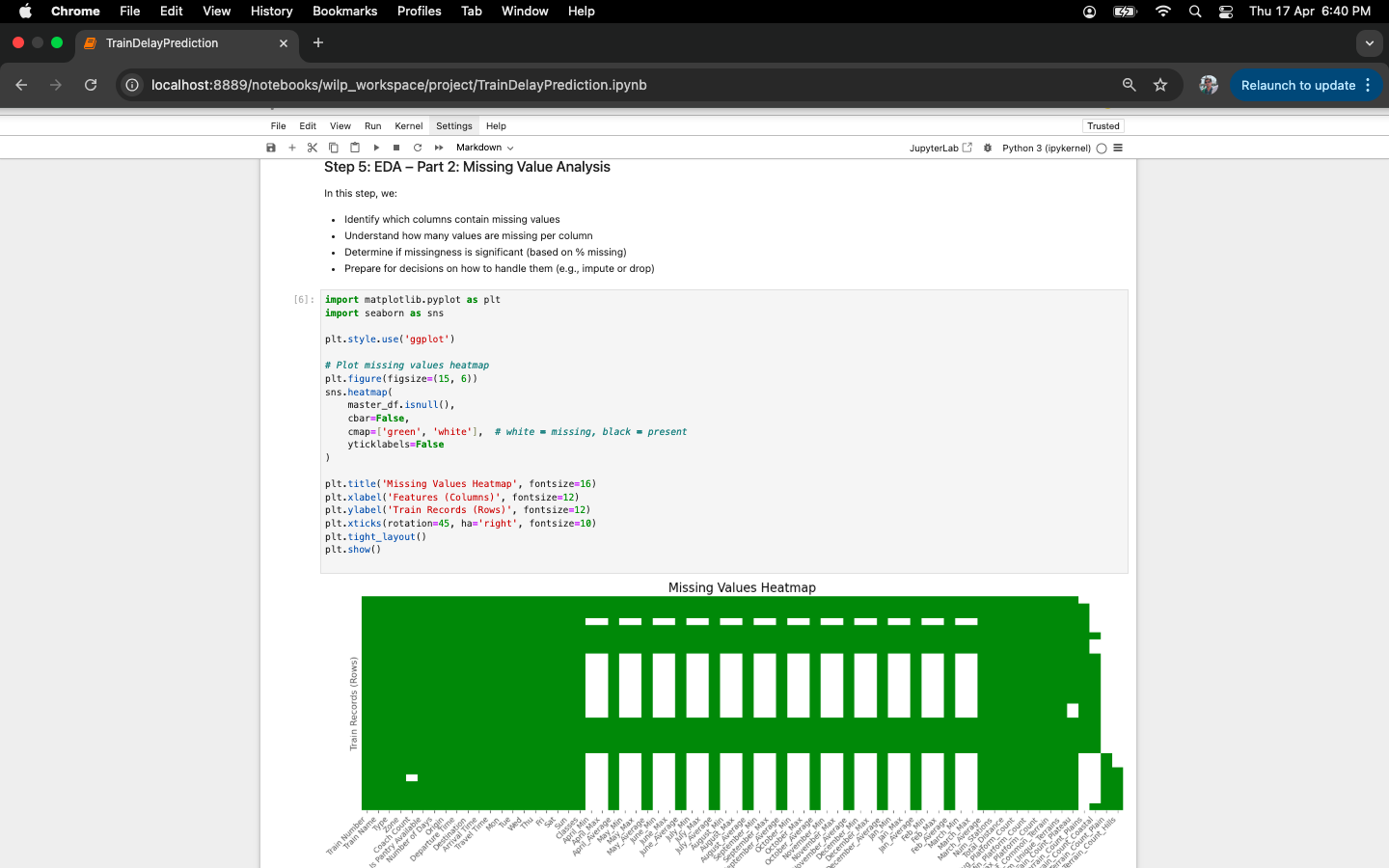
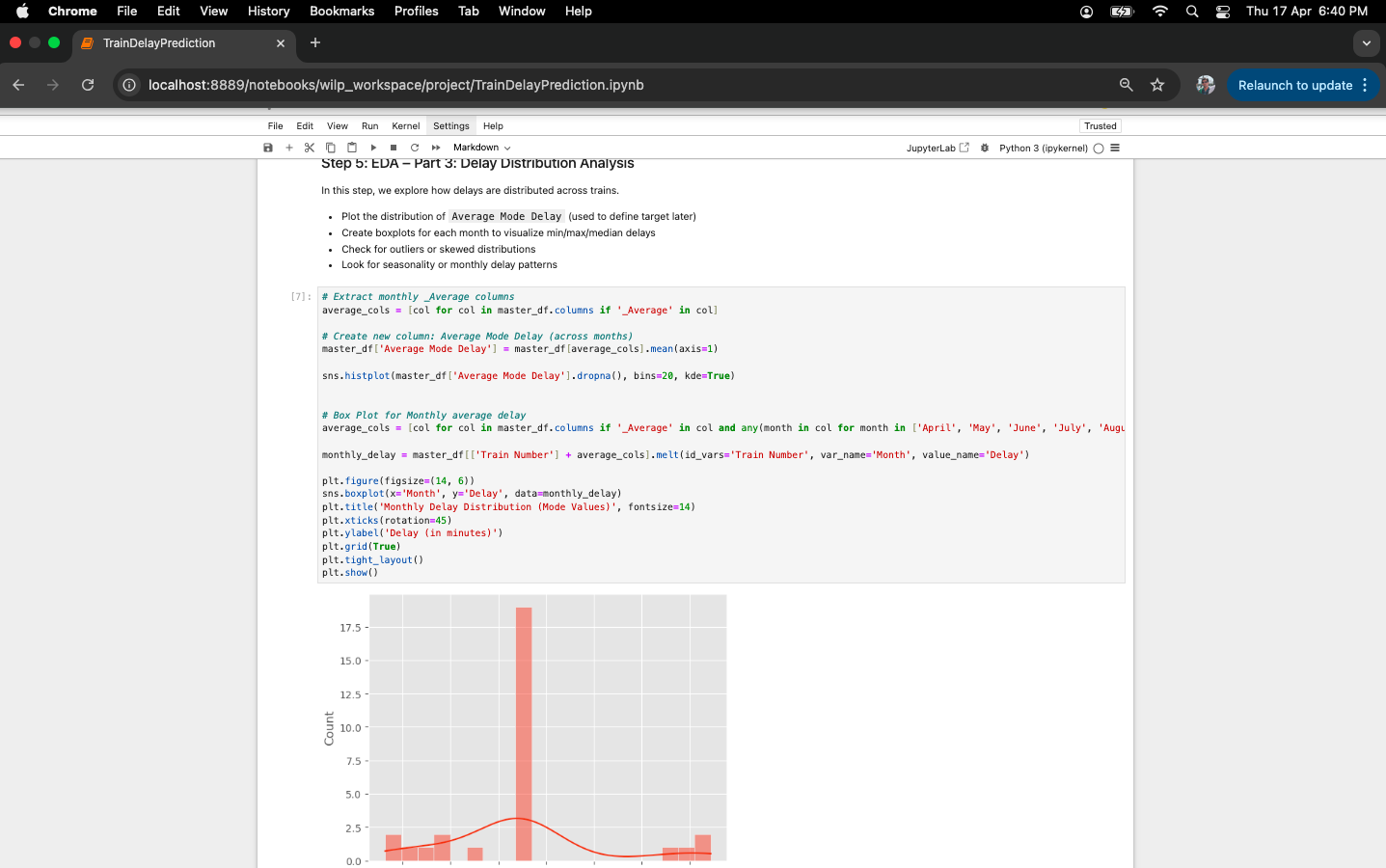
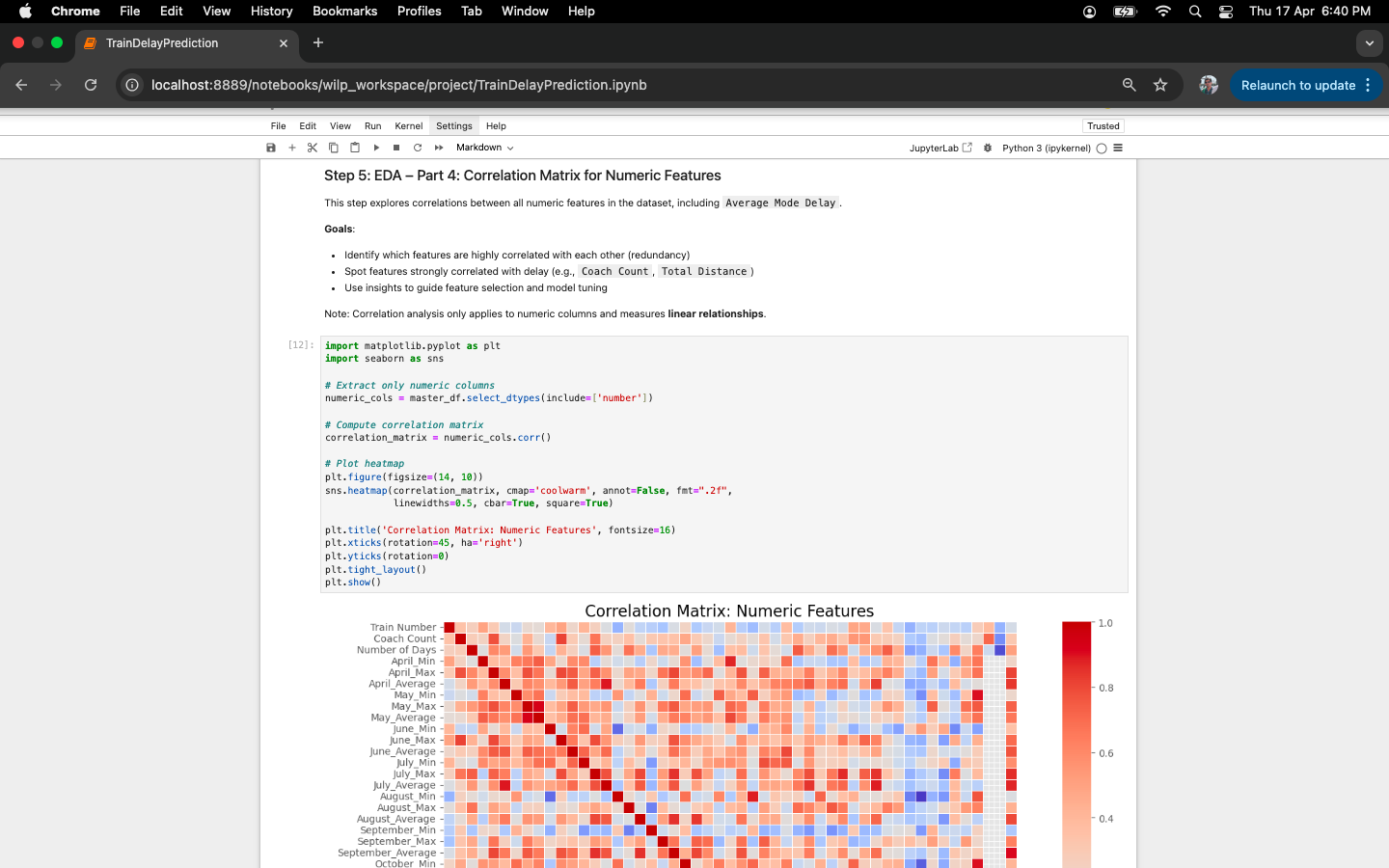
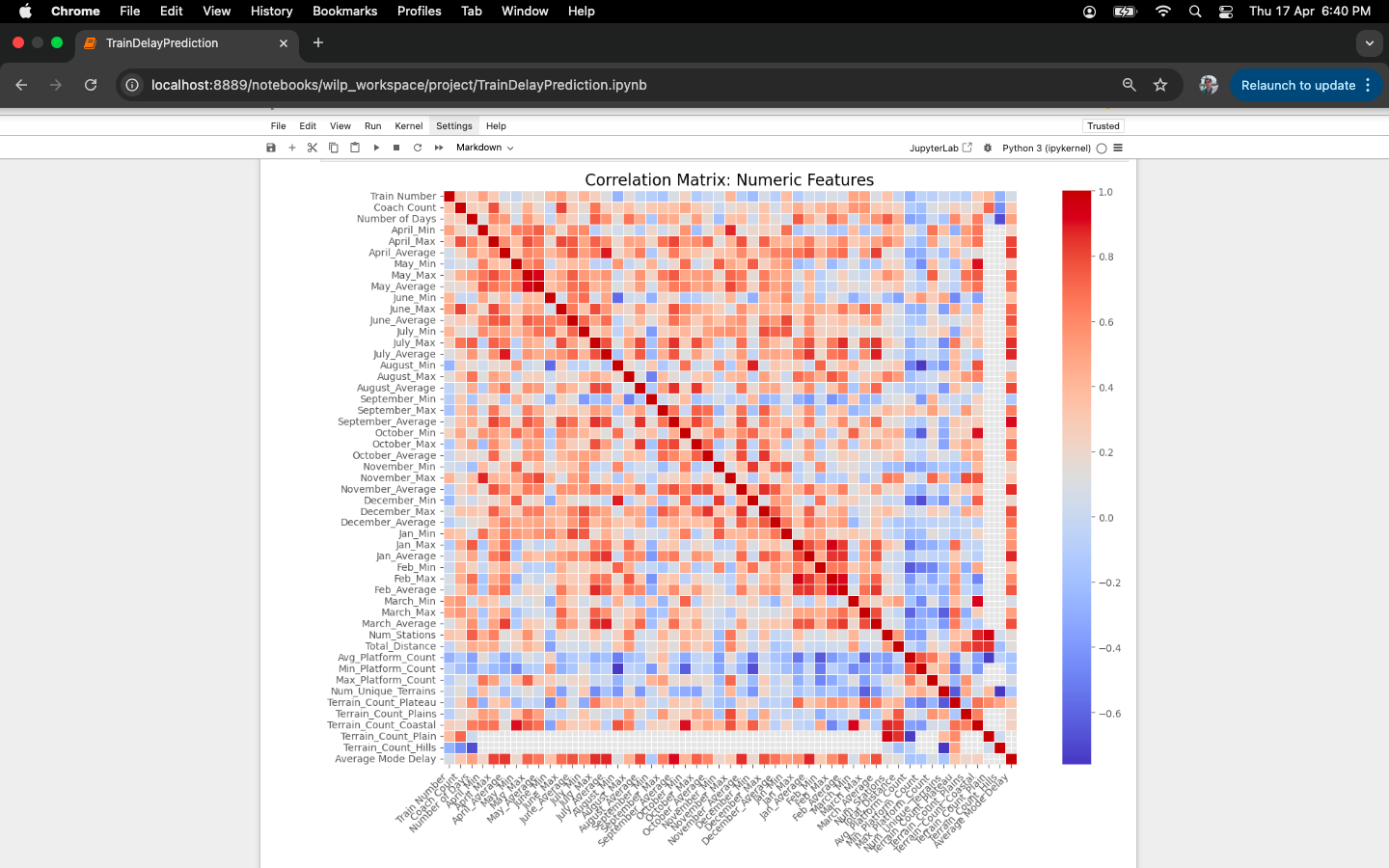
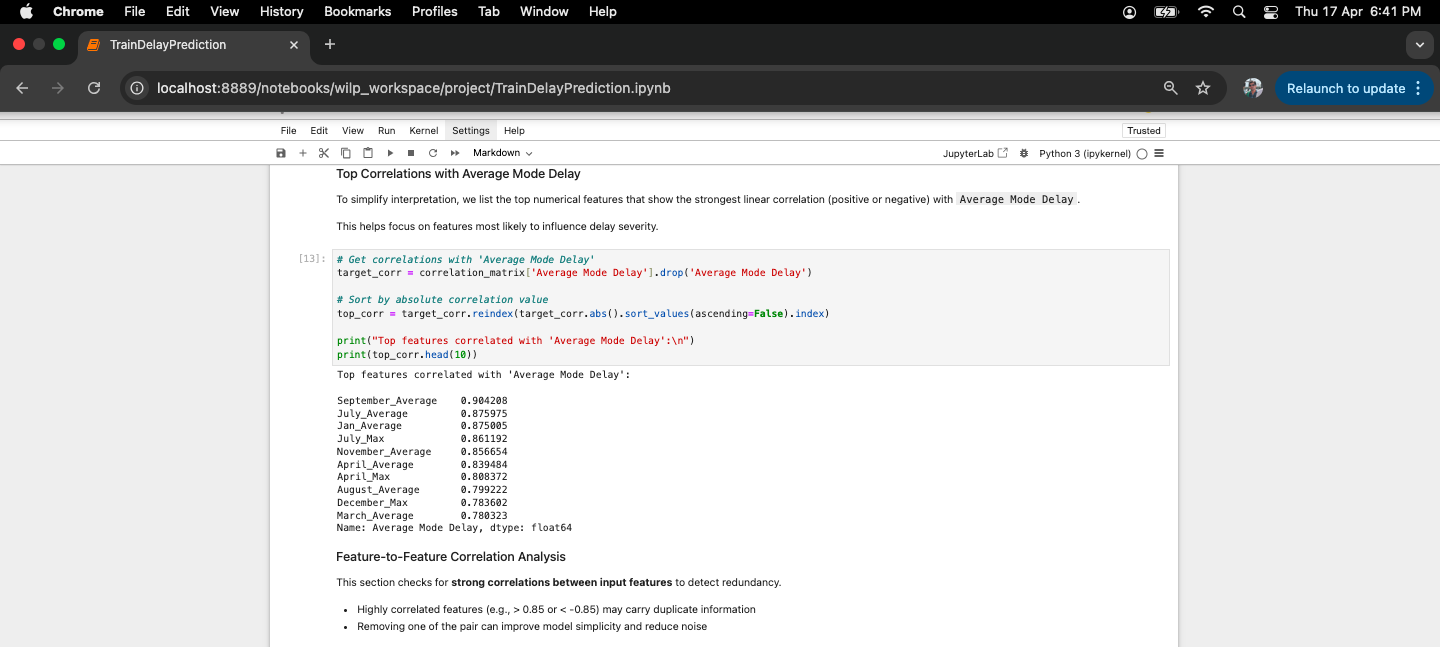
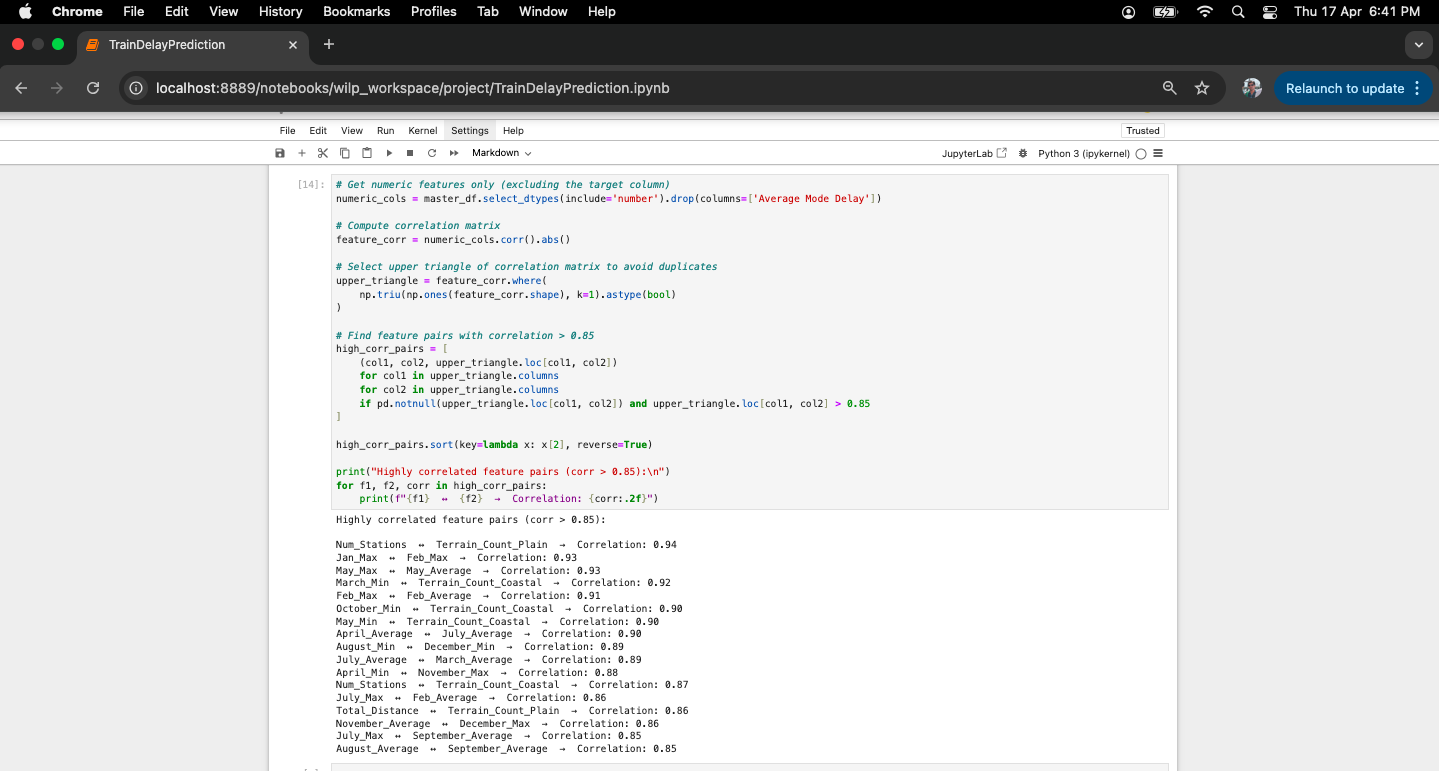
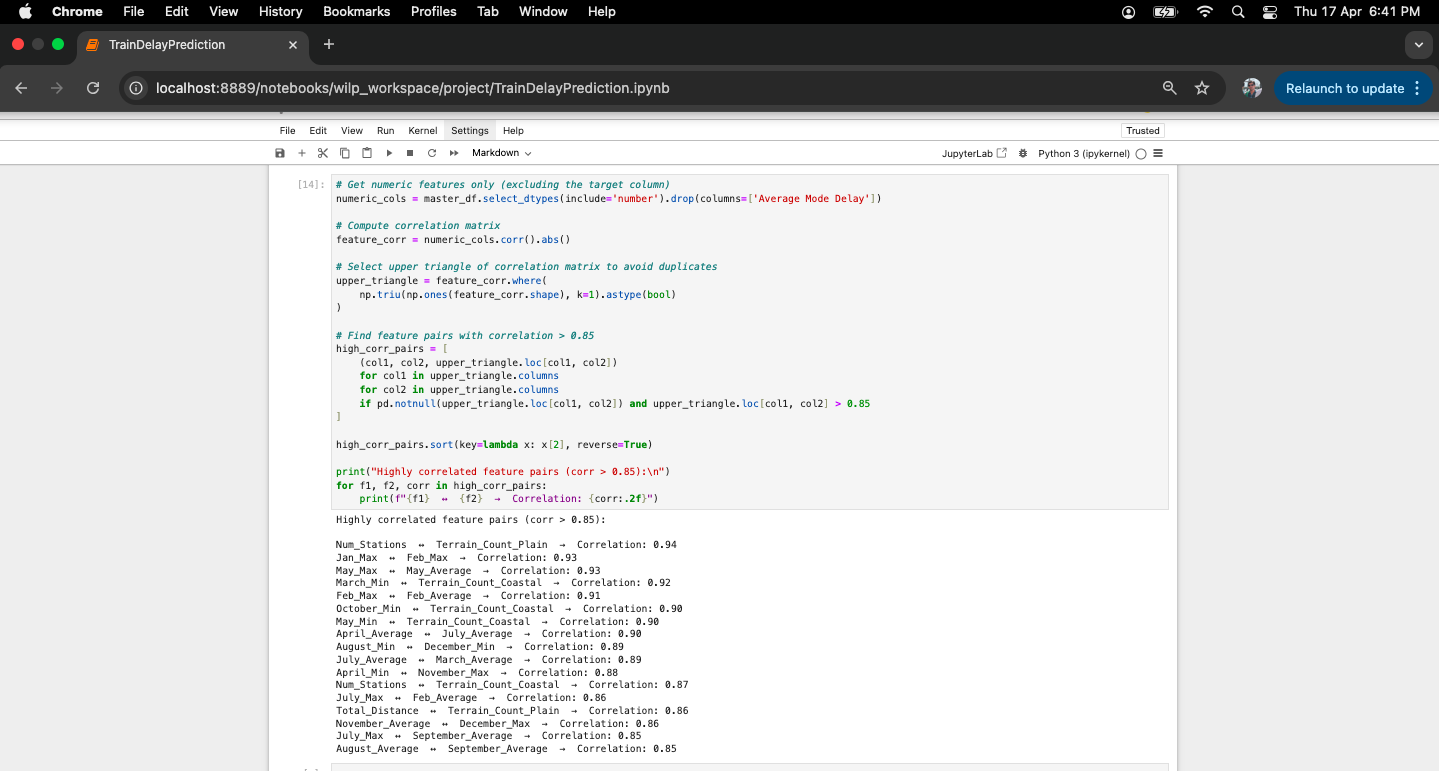
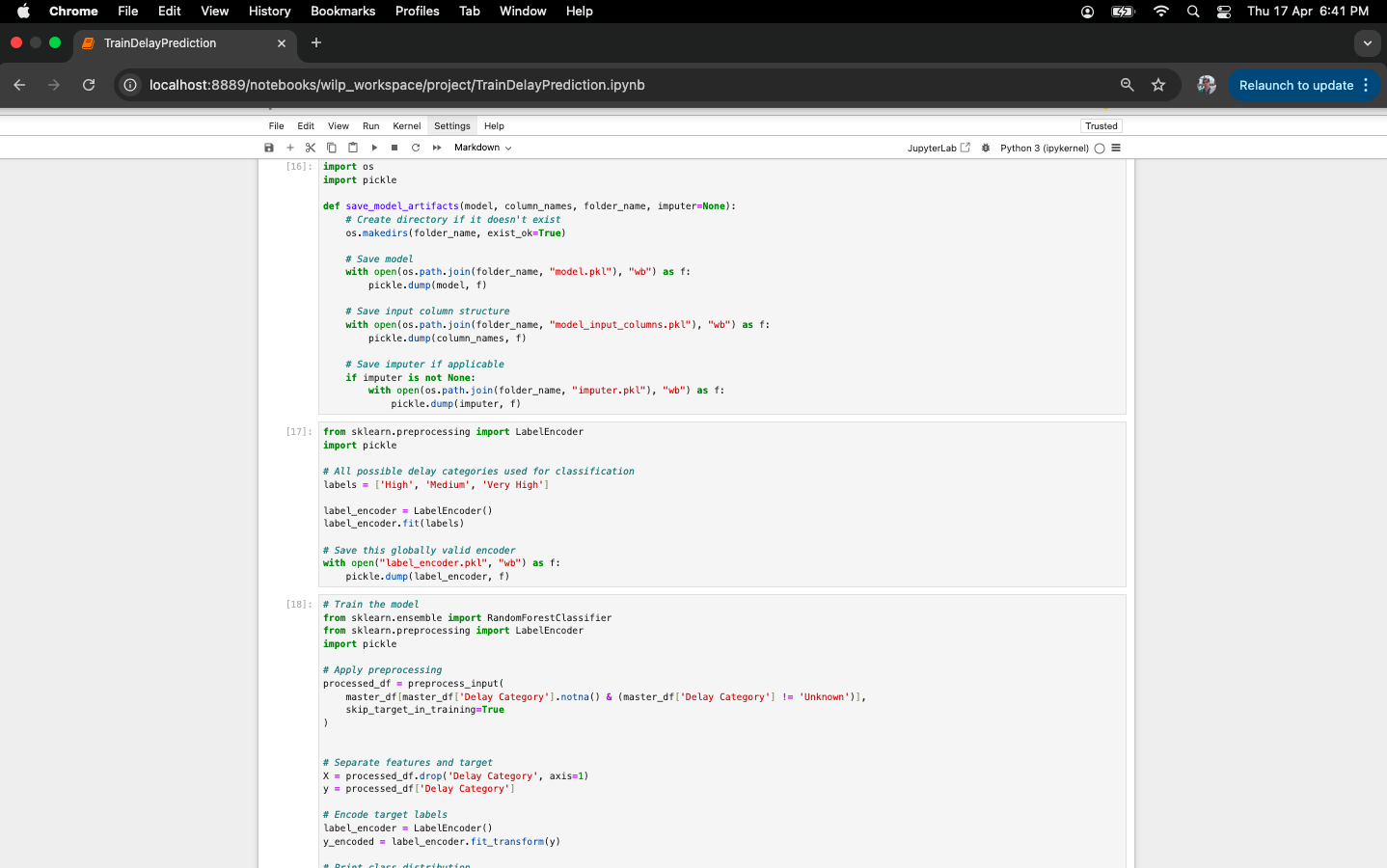
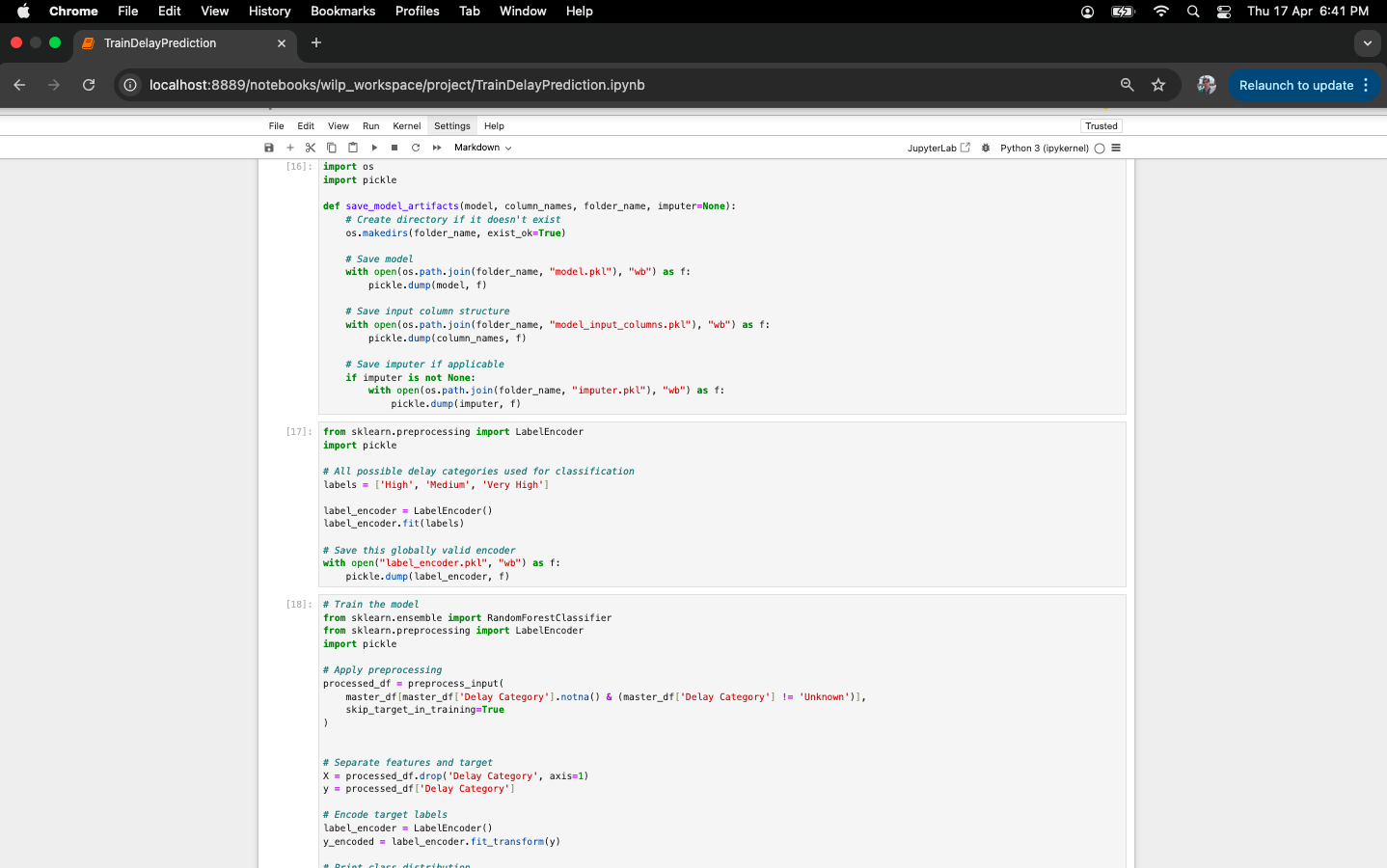
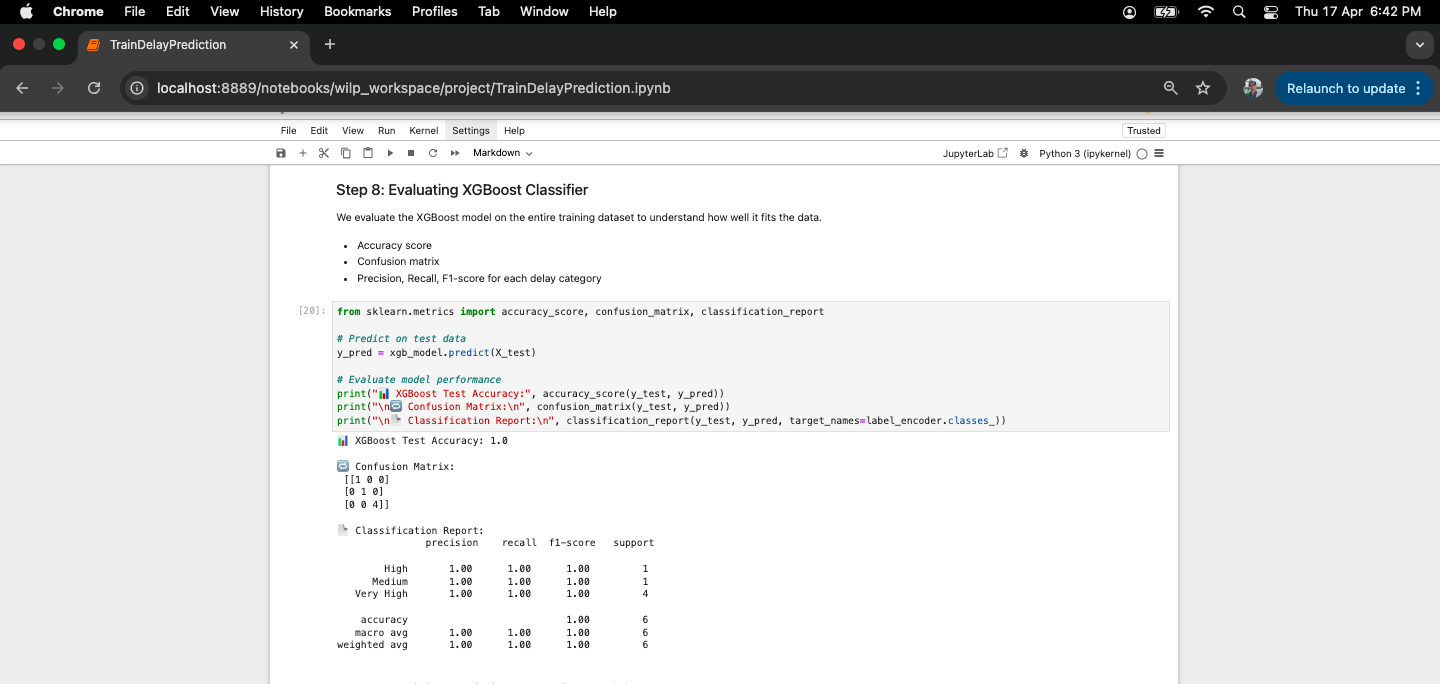
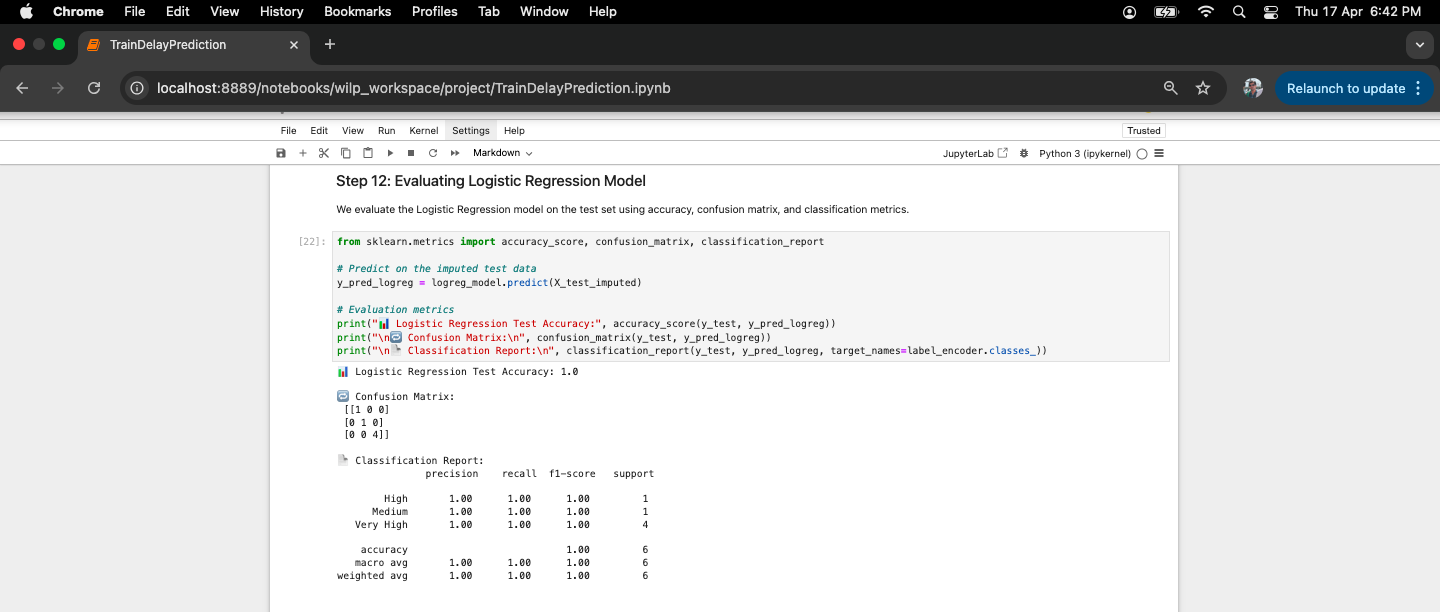
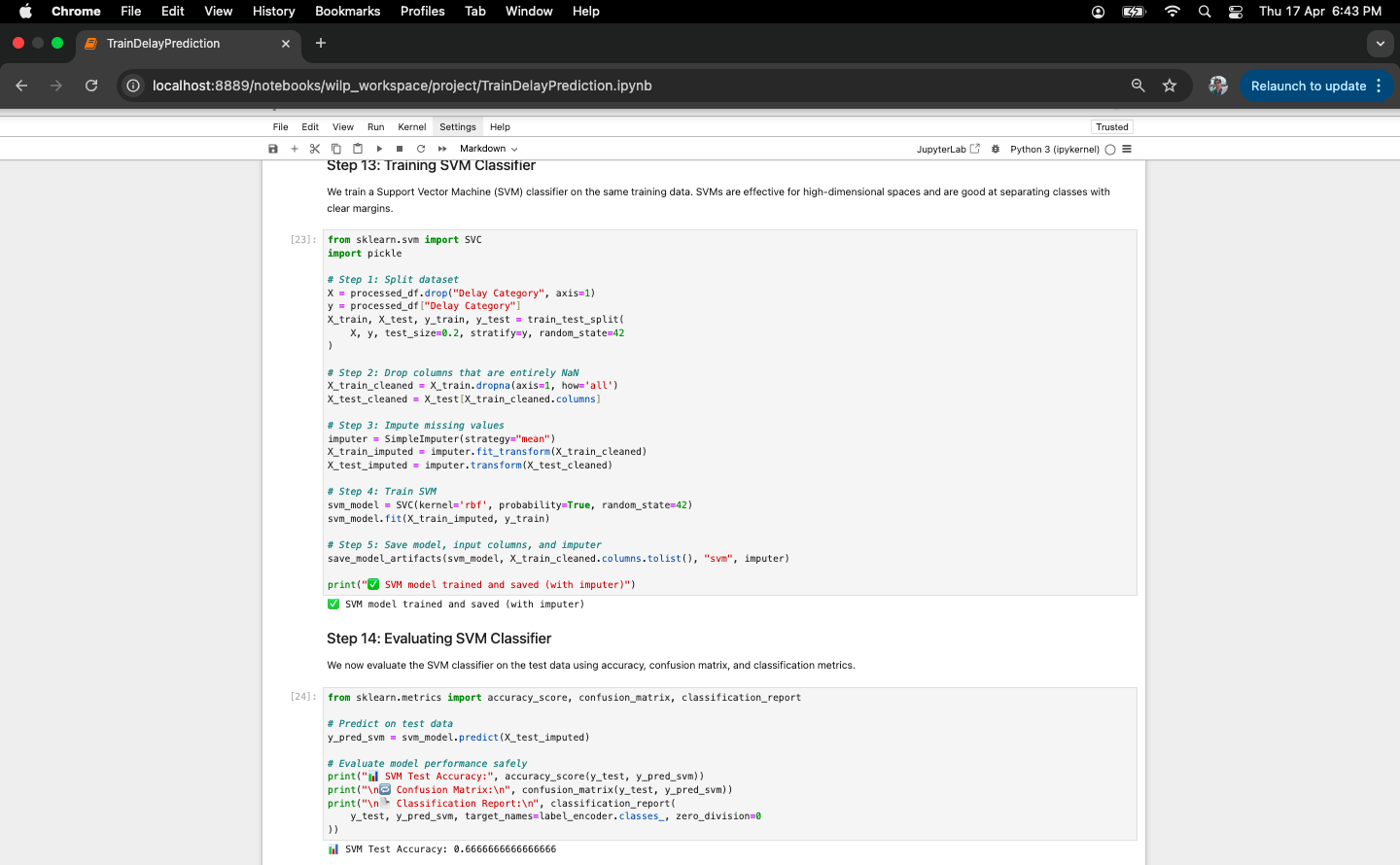
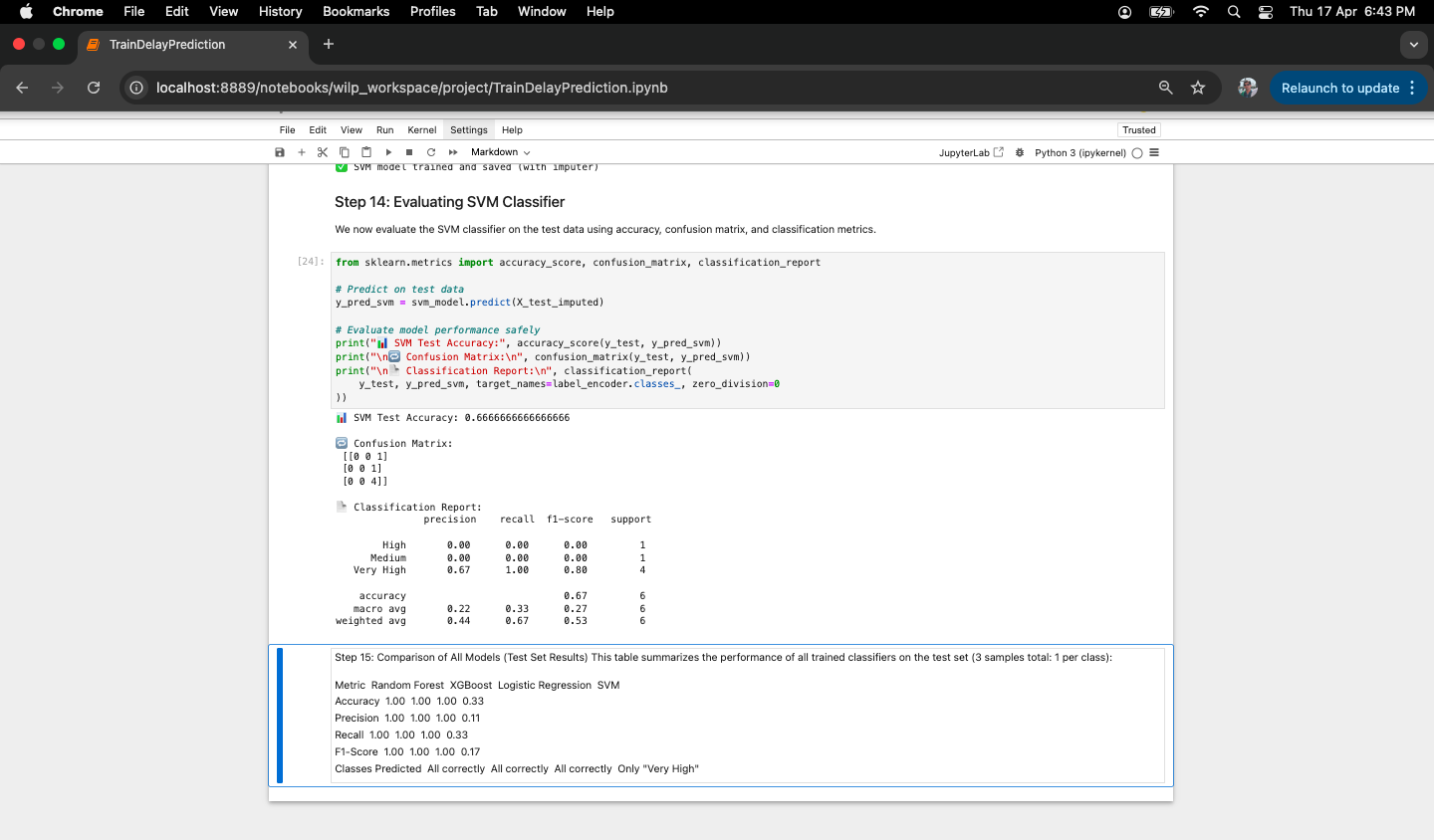
* Random Forest
* XGBoost
* Logistic Regression
* Support Vector Machine
* Decision Tree
* Neural Networks
* Naïve Bayes
* K-nearest neighbors

**Model Evaluation Summary:**



## 

## **Code & Screenshots**

1. Github repo : <https://github.com/ankitprjts/bitsGrp15Project>
2. Screenshots from Notebook:
   1. 
   2. 
   3. 
   4. 
   5. 
   6. 
   7. 
   8. 
   9. 
   10. 
   11. 
   12. 
   13. 
   14. 
   15. 
   16. 
   17. 
   18. 
   19. 
   20. 

## 

## **Deployment**

## **Objective:**

The deployment objective was to make the trained machine learning models accessible through an interactive and user-friendly web interface that allows users to input train details and get predicted delay categories in real time.

## **Tools and Technologies Used:**

1. Python: For building the entire pipeline including preprocessing, modeling, and inference logic
2. Streamlit: For developing the interactive web interface
3. Pickle: For saving and loading trained models, imputers, and encoders
4. Pandas & NumPy: For data manipulation
5. Scikit-learn & XGBoost: For training and predicting using ML models

## **Deployment Architecture Overview:**

The deployment architecture is designed to be modular and scalable. It uses a central Streamlit app (`app.py`) placed at the project root, and each trained model is stored in a dedicated directory. Each model directory includes the trained model file (`model.pkl`), input columns used for training (`model\_input\_columns.pkl`), and imputer (if needed).

When a user selects a model and submits input values, the app dynamically loads the corresponding model, applies the preprocessing logic (via `TrainDelayPrediction.py`), ensures that the feature columns match, applies imputation if required, and finally predicts the delay category.

## **Project Folder Structure:**

The following directory structure is used to organize the project:  
  
📁 project-root/  
│  
├── app.py  
├── label\_encoder.pkl  
├── TrainDelayPrediction.py  
├── random\_forest/  
│ ├── model.pkl  
│ └── model\_input\_columns.pkl  
├── xgboost/  
│ ├── model.pkl  
│ └── model\_input\_columns.pkl  
├── logistic\_regression/  
│ ├── model.pkl  
│ ├── model\_input\_columns.pkl  
│ └── imputer.pkl  
├── svm/  
 ├── model.pkl  
 ├── model\_input\_columns.pkl  
 └── imputer.pkl

## **Key Insights & Impactful/Important Features**

## **Key Insights**

1. The majority of trains experience low to moderate delays, with a few outliers that face very high delays.
2. Certain months show higher delay variability, indicating seasonal or operational patterns.
3. Routes with a higher number of stations and varied terrain are more prone to delays.
4. Trains with pantry facilities and higher coach counts do not necessarily correlate with higher delays, challenging some operational assumptions.
5. Some zones like NR and SCR exhibited higher variance in delays, possibly due to traffic congestion or regional infrastructure.

## **Important Features Identified**

1. Average Mode Delay: Aggregated target feature used for classification.
2. Travel Time: Derived from departure and arrival time; a critical feature for distinguishing long vs short routes.
3. Number of Stations: Higher values typically associate with more potential delays.
4. Platform Count (Avg/Min/Max): Proxy for infrastructure capacity; influenced route efficiency.
5. Train Type & Zone: Captures operational style and geographic spread; both show strong correlation with delays.
6. Departure and Arrival Time Bins: Helped capture time-of-day effects like peak hours or night-time buffers.
7. Terrain Categories: Presence of hills or coastal terrain introduces delay risk due to geographical constraints.

## **Future Work / Scope of improvements**

The current model for estimating train delays in Indian Railways, based on one year of historical data, provides valuable insights into delay patterns. However, there are several areas where the project can be further improved to increase its accuracy, scalability, and applicability. Below are some potential directions for future work:

#### **1. Expansion of Historical Data**

Currently, the model is trained on **one year of historical data** due to limitations in data access from the available websites (etrain.info and indianrail.gov). As there are no public APIs or access to data spanning more than a year, **future work** could focus on:

* **Exploring alternative data sources**: Working with Indian Railways directly or collaborating with research institutions for access to older data spanning **2–3 years**.
* **Using synthetic data** or **simulation techniques** to expand the training set and better account for long-term trends in train delays.

This will allow the model to capture **long-term seasonality**, **infrastructure changes**, and **systemic delays**, offering a more comprehensive view of the problem.

#### **2. Incorporating More Features**

The current model uses basic features like **Train details, route details, terrains, days of run etc.** Additional external factors that could be incorporated include:

* **Track maintenance schedules**: To understand how track repairs or construction might impact delays.
* **Previous day’s delay patterns**: Integrating delays from the **previous day** or historical data could help predict delays more accurately by capturing trends over time.
* **Passenger load and congestion**: Including data on passenger volume or train occupancy might offer insights into delays due to overcrowding or operational inefficiencies.

**3. Admin Dashboard for Ops**

Build a multi-tab Streamlit app that shows:

* Monthly delay trends
* Worst-affected routes
* Accuracy of models over time

**Impact:** Valuable for rail ops teams, not just passengers

**3. Smart Alerts / Notification System**

* Allow user to subscribe to a train ID and get predictions emailed/pushed before travel
* **Impact:** Improves user engagement and real-world utility

## **BIBLIOGRAPHY**

1. Reuters. (2025, January 8). *Indian Railways to spend over $22 billion on modernisation*. Reuters.<https://www.reuters.com/business/autos-transportation/indian-railways-spend-over-22-billion-modernisation-2025-01-08>
2. eTrain. (n.d.). *Train Running Status and Information Portal*. Data for the period April 2024 to March 2025 collected between March 3 and March 21, 2025, from<https://etrain.info/in>
3. Indian Railways. (n.d.). *Indian Railways Official Website*. Data for the period April 2024 to March 2025 collected between March 3 and March 21, 2025, from<https://indianrail.gov.in>
4. Gaurav, R. (n.d.). *Train Delay Estimation – Tutorial* [GitHub repository]. Retrieved March 3, 2025, from<https://github.com/R-Gaurav/train-delay-estimation/blob/master/doc/Tutorial.md>

## **Checklist of items for the Final report**

1. Is the Cover page in proper format? Y
2. Is the Title page in proper format? Y
3. Is the Certificate from the Mentor in proper format? Has it been signed? Y / N
4. Is Abstract included in the Report? Is it properly written? Y
5. Does the Table of Contents page include chapter page numbers? Y
6. Does the Report contain a summary of the literature survey? N
   1. Are the Pages numbered properly? Y
   2. Are the Figures numbered properly? Y
   3. Are the Tables numbered properly? Y
   4. Are the Captions for the Figures and Tables proper? Y
   5. Are the Appendices numbered? N
      1. Does the Report have Conclusion / Recommendations of the work? Y
      2. Are References/Bibliography given in the Report? Y / N
      3. Have the References been cited in the Report? Y / N
      4. Is the citation of References / Bibliography in proper format? Y / N